Natural Language Processing Methods for Detection of Influenza-Like Illness from Chief Complaints

Jia-Hao Hsu[1], Ting-Chia Weng[2], Chung-Hsien Wu[1] and Tzong-Shiann Ho[3]
[1] Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, TAIWAN
[2] Department of Family Medicine, National Cheng Kung University Hospital, Tainan, TAIWAN
[3] Departments of Pediatrics, National Cheng Kung University Hospital, College of Medicine, National Cheng Kung University, Tainan, TAIWAN
E-mail: {sky84003, tingchiaweng, chunghsienwu}@gmail.com, tsho@mail.ncku.edu.tw

***Abstract*— There are several existing studies on the application of medical chief complaints in disease classification. However, the lack of a standard vocabulary and high-quality interpretation of chief complaints hinder effective classification. This study uses a variety of methods to analyze chief complaints of preschool children to detect influenza-like illness. It is expected that a fast and effective tool can be designed to assist physicians in making diagnosis, and when facing a major outbreak, it can be quickly judged to control the outbreak as soon as possible. We use several natural language processing (NLP) technologies including deep learning methods, such as the currently popular BERT model, to classify Chinese chief complaints at emergency department to detect influenza-like illness. For model evaluation, the data in 2018 were used. The method based on BERT achieved the best accuracy of 72.87% for detection of influenza-like illness.**

I. INTRODUCTION

Influenza posed a significant burden of morbidity and mortality globally. The seasonal influenza caused about 3 million to 5 million severe cases worldwide, of which about 250,000 to 500,000 patients died annually [1]. Most cases of lethality occurred in the very young and the very old, or patients with underlying diseases [2]. The disease spectrum of influenza varied from mild or asymptomatic to severe inflammation with systemic involvement. A rapid detection of influenza from the early symptoms and signs remained challenging, especially among preschool children who could hardly express themselves properly and are difficult to evaluate in clinical practice. It is desirable to provide an appropriate tool for family and clinicians to improve the efficiency of early detection and to prevent the spreading of the disease.

This study focuses on preschool children aged 0-7 year-old with influenza-like illness (ILI). We applied the electronic medical records (EMR) to retrieve the chief complaints at triage station of emergency department to predict the clinical diagnosis of ILI based on ICD-9 codes at discharge. The aim of the study was to design a fast screening tool at low-cost to assist the detection of ILI among preschool children [3]. The goals were to train an auxiliary classification tool that can assist physicians in timely screening of ILI and to help predict the trends of disease spreading.

Early detection of ILI and timely prediction of the trends are important for public health. At the time of large-scale outbreak, such as the COVID-19 pandemic, the repeated waves of coronavirus overwhelmed hospitals in the threatened communities. Currently available rapid tests for influenza were still at high cost and yielded a low sensitivity and specificity. Polymerase chain reaction (PCR) provided a better sensitivity but it was time consuming and even more expensive. It is an urgent need to develop a reliable screening tool and timely surveillance system to prepare for the global pandemic.

Chief complaints from the patients were important clinical information that provided both subjective symptoms and objective signs [3-5]. At emergency department, chief complaints were recorded by a trained nurse at triage station. The information from chief complaints had been utilized to develop syndromic surveillance system for different diseases using natural language processing (NLP) [6-10] which is often used to deal with text-related tasks [11-13]. For example, rule-based bag-of-word (BOW) [14] matching techniques are most commonly used [10, 15]. Researchers defined their own or use the *tf-idf* score [16] of each word in the data as a basis to define the content of the bag of words [6], and to search for the matching bag of words from each chief complaint to classify the diseases. The deep learning NLP technology was also gradually paid attention to the classification of diseases. The main role of the NLP technologies in this work was semantic understanding. For each word, this study considered the contextual relationship using different models for word embedding processing, such as word2vev [17, 18], fastText [19], GloVe [20] and other basic word embedding models. These methods could give each word a fixed representation as the feature of the word.

With the rapid evolution of a large number of training corpus and large-scale deep models, strong and effective language models such as ELMO [21] and BERT [22] have emerged, which have more complete and effective processing and performance for word embedding and semantic understanding. Then, the word embeddings are input into the classifier as the characteristics of each data for disease detection.

Some medical complaints related studies also used multimodal data to classify diseases [8, 23, 24], such as patients' personal information, vital signs or medical images. These studies extracted features from each piece of information, and used a variety of different fusion methods to properly mix the features to achieve modal mutual benefit.

To the best of our knowledge, there were some studies used chief complaints to develop syndromic surveillance system for influenza-like illness. This study also referred to the above

literatures for the analysis of ILI from the data collected at emergency department of National Cheng Kung University Hospital (NCKUH) from Jan 2015 to Dec 2018 [25]. Different from the above literature, we also referred to the judgment of professional physicians on the clinical pathology of influenza-like illness as a classification consideration and tried to mix the information obtained in the chief complaint with other physiological information to obtain a better accuracy.

II. METHODS

For comparison, the methods can be roughly divided into two categories: non-deep learning mechanism and deep learning mechanism. In the method of non-deep learning mechanism, we first use the aforementioned bag-of-word matching technique as the most preliminary experimental method, and then use a linear regression model as the classifier. The method of deep learning mechanism uses the retrained word embedding model and the pre-trained BERT as the feature extraction model of the chief complaint and puts the extracted representation to the fully connected layer for classification. The details are described below.

A. Data Pre-Processing

First of all, we deal with all the main complaints through the stuttering system. Then delete the redundant words that may appear in the main complaint and the indefinite auxiliary words carried by the nursing staff during the recording, such as the words "Enter" or "transfer". In addition, the English words in the input sentence are also converted to a unified lowercase format, so as to avoid converting these non-pre-trained English words into unknown token (UNK) in the BERT pre-processing step, which may cause subsequent judgments.

B. Rule-Spam Method

The Rule-spam method [26] uses the words in the bag of words as the basis for judgment. The method of selecting the words in the bag of words uses the common method of calculating the word frequency score in NLP to find the words with a higher *tf-idf* score from the data, as shown in the following equations.

$$score_t = tf_{t,ILI} \times idf_t - tf_{t,nonILI} \times idf_t \quad (1)$$

$$tf_{t,d} = \frac{n_{t,d}}{\sum_{k=1}^{T} n_{k,d}} \quad (2)$$

$$idf_t = \log\left(\frac{D}{d_t}\right) \quad (3)$$

We apply the spam score mechanism to classify each word. Then we divide the original data into two documents. The data in Document 1 are the chief complaints marked as ILI document, and the data in Document 2 are the chief complaints marked as non-ILI document. We calculate the scores of *tf-idf* in these two documents separately and calculate the score of each word in the two categories. Among them, $tf_{t,d}$ represents the word frequency of the *t*-th word appearing in document *d*, which is used to indicate the characteristics of the words in each document. *Idf* is the inverse document frequency. It is used to express the characteristics of each word in all documents. From equation 3, it can be seen that if a word appears in more documents, the corresponding *idf* value will be smaller. This means that the word may contain less information and is less important. Subtract the non-ILI word score from the ILI word score to get the spam score for that word. It calculates the importance of the word in the two documents. If a word has high score in the ILI document and the non-ILI document, it is not an important keyword for judging ILI. If a word has a high score in the ILI document, but a low score in the non-ILI document, the spam score obtained after subtraction will be high. It means that it is an important keyword for judging ILI. We select the top *n* words with high spam score from the experiment as the judgment word of the Rule-spam method. If the input complaint contains the words in the Rule-spam bag, it is judged as ILI, otherwise it is not.

C. Rule-Best Method

Rule-best method also uses BOW matching technology, but the bag is different from the Rule-spam method. We use the nine major syndrome-groups of respiratory tract infections provided by professional physicians as the content of the BOW, as shown in Table 1.

TABLE. 1
NINE MAJOR SYNDROME-GROUPS OF RESPIRATORY TRACT INFECTIONS

Group	keywords
01-Fever	Fever, heat, chillness, hot, sweating, chill, sweating
02-Pain	Pain
03-Upper respiratory tract	Stuffy nose, rhinorrhea, cough, throat
04-Lower respiratory tract	Sputum, discharge, dyspnea, respiratory distress
05-Eye	Eye
06-Ear	Ear
07-Skin	Rash
08-Gastrointestinal	Nausea, vomiting, diarrhea, abdomen unspecified, milk, jaundice
09-Activity	Appetite, activity, energy, drowsy, fatigue, sickness, irritable, cry

Then the method matches each word in each input sentence with the words in the nine major word bags. If the matching is successful and the input (complaint) is judged as an ILI patient, otherwise it is judged as a non-ILI patient. Rule-best method selects the best combination of BOW from these nine bags as the basis for judgment, as shown in equation 4.

$$\text{Rule}_{\text{best}} = arg\max_{s}(Acc(s)), s \in S \quad (4)$$

We use all 510 ($\sum_{i=1}^{9} C_i^9$) BOW combinations as the basis for judging ILI, and calculate the accuracy (Acc) for each combination. From these combinations, the BOW combination with the highest accuracy is selected as the Rule-best combination.

D. Logistic Regression Method

Before using this method, the data must be organized into vector (feature) and target formats. We convert the chief

complaint sentence into a vector form. The method of conversion is to compare whether each chief complaint sentence contains the words in the nine major word bags using a rule-based approach. In this way, each chief complaint sentence can be projected into a vector with a dimension of 9, as shown in Equation 5.

$$rule\ vector_i^k = \begin{cases} 0, \forall\ word \in k, but\ word\ \notin\ group\ i \\ 1, \exists\ word \in k, and\ word \in group\ i \end{cases} \quad (5)$$

In the *k*-th chief complaint, if a word that also belongs to the *i*-th bag of the nine major groups can be found, then the *i*-th dimension of the rule-vector of this chief complaint is 1, or 0 if not. For example, the main complaint is: "The patient came to the clinic for cough, vomiting", where "cough" is included in the third group "upper respiratory tract", and "vomiting" is included in the eighth group "gastrointestinal", so this chief complaint can be projected as a vector of <0,0,1,0,0,0,0,1,0>. We convert all chief complaints into the rule-vector form, and let the model learn a linear regression line with these input vectors [27]. The target end and the non-target end are classified at both ends of the regression line. The trained regression model will input the test data and project it in this regression space, and get its bias toward the target or non-target in this space (or called its probability value that can be judged as the target).

E. GloVe Method

We also use machine learning to do experiments on influenza-like illness prediction. The first deep learning model is to use the word embedding model GloVe. Considering that the medical chief complaint is different from the general dialogue content, we do not use the pre-trained model, but use our corpus to do a complete training of the model, hoping that the trained word embedding is suitable to characterize the medical chief complaint. The GloVe model mainly generates a fixed representation and produces the same word embedding for the same word. The training process will refer to the relationship between the context of each word, as shown in Equation 6.

$$w_i^T \bar{w}_j + b_i + \bar{b}_j = log(X_{ij}) \quad (6)$$

Each element in the matrix *X* represents the number of times the word *i* and the context word *j* appear together in a context window of a certain size. w_i^T and $\bar{w}_j$ are the final word vectors required by the model, and b_i and $\bar{b}_j$ are the offsets of the two word vectors, respectively, which are obtained by training steps.

F. BERT Method

We use a standard 12-layer BERT-BASE model that has been pre-trained with a large amount of Chinese corpus. The method performs word vector compression on the main complaint sentence in Chinese. Among them, BERT's self-attention mechanism will disseminate all token information to other tokens, so the start token can contain the information of all input sentences, and it is often used in the subsequent research of the BERT model. Therefore, we use the compressed start token (CLS) to project the start token information into the ILI classification through a layer of all-linked hierarchy.

III. Experimental Results

A. Dataset

The data used in this study was the chief complaints at emergency department of NCKUH. There were a total of 61,208 records in the data, with patients ranging from 0 to 35 years old. Each patient record has 135 items including gender, age, diagnosis code, chief complaint and other vital signs. We first filter the data by two steps. The first step is to delete the data of patients older than 7 years old. Because we believe that the method of obtaining the chief complaints should be unified, and some patients older than 7 years old may provide information through self-reports, and preschool children under 7 years old usually provide information through family-reports. The second step is to delete the patient information that repeats the visit within 24 hours. This step is intended to reduce the recurrence of specific data. In the end we filtered out 44,658 available records.

There are no labels of ILI in the original data, which contain only the diagnosis code of the doctor's initial diagnosis. We use the diagnostic code marked by the doctor to label ILI or not. We used the definition table of influenza-like diagnosis codes proposed by [25]. If the diagnosis code marked by the doctor matches the diagnosis code in the table, the data will be marked as ILI patients, otherwise, non-ILI will be marked. The filtered data is shown in Table2.

TABLE. 2
DATA STATISTIC

	ILI		Non-ILI		Total
Emergency Visits	20237		24421		44658
Girl	9152	45%	10618	43%	19770
Boy	11085	55%	13803	57%	24888
Age-Group					
Infant (<1y)	3914	19%	5643	23%	9557
Toddler(1~2y)	8632	43%	9165	38%	17797
Preschooler(3~7y)	7691	38%	9613	39%	17304

B. Performance Evaluation

We used a total of 33,765 records from 2015 to 2017 in the filtered data as the training data, and 10,893 records in 2018 as the testing data. The experiment was conducted in the

TABLE. 3
PERFORMANCE ON ONLY CHIEF COMPLAINTS

Methods	Accuracy	Sensitivity	Specificity	PPV	NPV
Rule-Spam	69.76%	86.51	56.00	61.77	83.47
Rule-Best	70.46%	85.98	57.70	62.56	83.35
Regression	71.33%	77.82	66.00	65.29	78.36
GloVe	71.01%	70.66	66.68	75.06	71.01
Bert	72.87%	66.06	78.47	71.61	73.77

aforementioned methods, and the results obtained using only the chief complaint as input are shown in Table 3.

The combination of Rule-best method achieving the best accuracy is the combination {01,06}. The BOW to get the best accuracy in Rule-spam is a bag using the five words: {fever, chills, 38 degrees, sickness, mobility}. Rule-based methods simply check whether there is a word in the bag of words in the chief complaint. It is likely to cause over-judgment, which leads to the phenomenon with the lowest accuracy but high sensitivity. The regression model considers more information about the syndromes that are not included in the nine syndrome-groups than the rule-based method, so the resulting efficiency is higher than the rule-based method. The method of deep learning is different from the previous methods (only keywords are considered). It also considered the severity of the symptoms and other factors, so it obtained the highest accuracy. However, there is no significant difference in the results of the various methods. We believe that the chief complaint of the influenza-like illness identified by the doctor may also rely on the keywords of the nine syndrome-groups, so the rule-based method was only 3% less accurate than the deep learning method.

C. *Predicted Trends*

We present the testing data in units of weeks, and record the number of patients predicted to be ILI by each method each week, and present them in the following figure. The red line is the week number of clinically diagnosed ILI, which can be regarded as the ground truth.

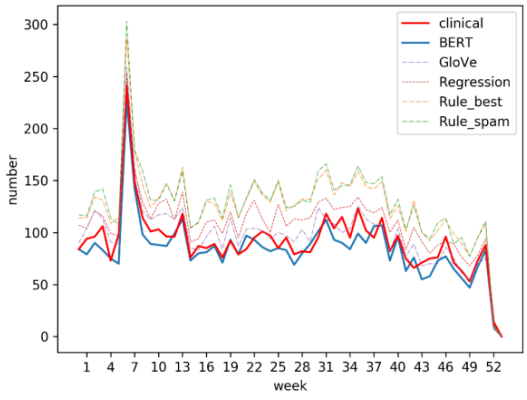

Fig. 1 Predicted Trends in 2018 data

Figure 1 shows the prediction amount of each method for patients throughout the year and it verifies that the aforementioned rule-based methods over-judge patients as ILI. It can also be seen that the prediction of BERT in quantity is most consistent with the ground truth. Over-prediction methods can increase the sensitivity of epidemic prevention, but at the same time may also cause waste of medical resources. It can be seen from the figure that winter vacation is an obvious peak period of influenza. If the model considers the time factor, it may improve the prediction accuracy.

D. *ROC Curve*

Receiver operating characteristic (ROC) curve is a curve commonly used to analyze the binary classifiers. The classifiers give all testing data the classification probabilities. Generally, the threshold value of 0.5 is used. The method checks whether the probability value obtained from each test data is greater than the threshold value. If it exceeds, the judgment value is 1 and the none value is 0. By adjusting the threshold, the specificity obtained by the model under different sensitivities can be obtained and plotted as the ROC curve as shown in Figure 2. We also draw the results from other rule-based methods in the figure, but the rule-based method does not output the classification probability value as it cannot be drawn into a curve by adjusting the threshold. Rule-01 represents using the "fever" group of the nine syndrome-groups as the basis for judgment.

The evaluation method of ROC curve is to calculate the area under the curve (AUC). In Figure 2, the AUC of GloVe is 76.77,

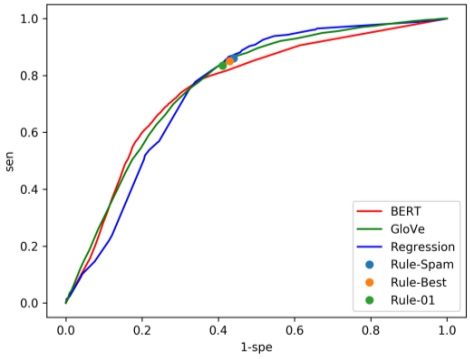

Fig. 2 ROC Curve for The Methods. Sen: Sensitivity; Spe: Specificity

BERT is 75.54 and regression is 75.26. It can be seen that BERT and GloVe are still slightly better than regression method. Probably because the chief complaint may be different from the general corpus, GloVe trained using our data has better AUC than the pre-trained BERT. It can also be seen in the figure that the positions of Rule-spam, Rule-best, and Rule-01 are all above the BERT curve. Although these methods cannot be shown to be better than BERT, they can be seen that these methods have a certain classification effect. All three methods include the word "fever" in their BOW, so it is inferred that "fever" may be a key basis for the judgment of ILI.

IV. CONCLUSIONS

With this research, we analyzed several methods applied to the classification of influenza-like illness in pediatric patients. The model of BERT achieved the best accuracy at 72.87%. The methods could provide timely and easy-to-operate diagnostic tools at clinical settings, which could assist physicians in making initial diagnose. In the future, we will add other clinical information together for analysis, such as the time of visit and other vital signs. However, there are still some unresolved problems or limitations in this study, such as the missing data and the significant impact of the differences of the recorders. If each recorder has different habits or different professional terms, it will affect the accuracy of the classifier. These are issues that need to be improved in the future.

REFERENCES

[1] V. N. Petrova and C. A. Russell, "The evolution of seasonal influenza viruses," *Nature Reviews Microbiology,* vol. 16, no. 1, p. 47, 2018.

[2] L. A. Grohskopf, E. Alyanak, K. R. Broder, E. B. Walter, A. M. Fry, and D. B. Jernigan, "Prevention and control of seasonal influenza with vaccines: recommendations of the Advisory Committee on Immunization Practices—United States, 2019–20 influenza season," *MMWR Recommendations and reports,* vol. 68, no. 3, p. 1, 2019.

[3] S. H. Lee, D. Levin, P. D. Finley, and C. M. Heilig, "Chief complaint classification with recurrent neural networks," *Journal of biomedical informatics,* vol. 93, p. 103158, 2019.

[4] P. Manu, D. A. Matthews, and T. J. Lane, "The mental health of patients with a chief complaint of chronic fatigue: a prospective evaluation and follow-up," *Archives of Internal Medicine,* vol. 148, no. 10, pp. 2213-2217, 1988.

[5] A. J. Beitel, K. L. Olson, B. Y. Reis, and K. D. Mandl, "Use of emergency department chief complaint and diagnostic codes for identifying respiratory illness in a pediatric population," *Pediatric emergency care,* vol. 2, no. 6, pp. 355-360, 2004.

[6] G. Li, H. Song, H.-N. Liang, Y. Qu, L. Liu, and X. Bai, "Medical Diagnosis by Complaints of Patients and Machine Learning," in *2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 2019: IEEE, pp. 1-5.

[7] H. Song, G. Li, Z. Liu, and X. Bai, "Using Structured event to represent complaints of patients: a medical assistant for doctors," in *2019 IEEE 5th International Conference on Computer and Communications (ICCC)*, 2019: IEEE, pp. 2193-2197.

[8] Q. Wang, D. Yang, Z. Li, X. Zhang, and C. Liu, "Deep Regression via Multi-Channel Multi-Modal Learning for Pneumonia Screening," *IEEE Access,* vol. 8, pp. 78530-78541, 2020.

[9] S. Mulyana, S. Hartati, and R. Wardoyo, "A Processing Model Using Natural Language Processing (NLP) For Narrative Text Of Medical Record For Producing Symptoms Of Mental Disorders," in *2019 Fourth International Conference on Informatics and Computing (ICIC)*, 2019: IEEE, pp. 1-6.

[10] G. Veena, R. Hemanth, and J. Hareesh, "Relation Extraction in Clinical Text using NLP Based Regular Expressions," in *2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, 2019, vol. 1: IEEE, pp. 1278-1282.

[11] M.-H. Su, C.-H. Wu, and Y. Chang, "Follow-Up Question Generation Using Neural Tensor Network-Based Domain Ontology Population in an Interview Coaching System," in *INTERSPEECH*, 2019, pp. 4185-4189.

[12] M.-H. Su, C.-H. Wu, K.-Y. Huang, and Q.-B. Hong, "LSTM-based text emotion recognition using semantic and emotional word vectors," in *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, 2018: IEEE, pp. 1-6.

[13] M.-H. Su, T.-H. Yang, W.-H. Lin, and C.-H. Wu, "Answer segmentation for question answering using latent dirichlet allocation and delta Bayesian information criterion," in *2016 International Conference on Orange Technologies (ICOT)*, 2016: IEEE, pp. 9-12.

[14] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: a statistical framework," *International Journal of Machine Learning and Cybernetics,* vol. 1, no. 1-4, pp. 43-52, 2010.

[15] L. Wang, Q. Xu, and S. Li, "Utility Balanced Classification for Automatic Electronic Medical Record Analysis," in *2018 5th International Conference on Systems and Informatics (ICSAI)*, 2018: IEEE, pp. 1093-1098.

[16] J. Ramos, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, 2003, vol. 242: New Jersey, USA, pp. 133-142.

[17] C. McCormick, "Word2vec tutorial-the skip-gram model," ed: Retrieved, 2016.

[18] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781,* 2013.

[19] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, "Fasttext. zip: Compressing text classification models," *arXiv preprint arXiv:1612.03651,* 2016.

[20] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532-1543.

[21] M. E. Peters *et al.*, "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365,* 2018.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805,* 2018.

[23] M. Ozkan, B. M. Dawant, and R. J. Maciunas, "Neural-network-based segmentation of multi-modal medical images: a comparative and prospective study," *IEEE transactions on Medical Imaging,* vol. 12, no. 3, pp. 534-544, 1993.

[24] F. Liu, L. Zhou, C. Shen, and J. Yin, "Multiple kernel learning in the primal for multimodal Alzheimer's disease classification," *IEEE journal of biomedical and health informatics,* vol. 18, no. 3, pp. 984-990, 2013.

[25] T.-C. Weng, H.-Y. R. Chiu, S.-Y. Chen, F.-Y. Shih, C.-C. King, and C.-C. Fang, "National retrospective cohort study to identify age-specific fatality risks of comorbidities among hospitalised patients with influenza-like illness in Taiwan," *BMJ open,* vol. 9, no. 6, p. e025276, 2019.

[26] M. Sasaki and H. Shinnou, "Spam detection using text clustering," in *2005 International Conference on Cyberworlds (CW'05)*, 2005: IEEE, pp. 4 pp.-319.

[27] D. G. Kleinbaum, K. Dietz, M. Gail, M. Klein, and M. Klein, *Logistic regression*. Springer, 2002.